DOCUMENT RESUME

ED 091 398 TM 003 591

AUTHOR Sockloff, Alan L.
TITLE The Effect of Pooling Two Heterogeneous Subgroups on the Product-Moment Correlation Coefficient.
PUB DATE [Apr 74]
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Analysis of Variance; *Correlation; Hypothesis Testing; *Mathematical Models; *Research Problems; *Statistical Analysis; Tests of Significance

ABSTRACT

An equation was derived to determine the relationship between the pooled within-subgroup r (correlation coefficient) and the r obtained from the total group data. It was, thus, possible to assess the amount of distortion introduced by pooling heterogeneous subgroups. As a basis for deciding whether to pool two subgroups in order to calculate a single r for the total group, a two-stage procedure was recommended: (1) comparison of the two within-subgroups r's; and (2) comparison of the total group r and the pooled within-subgroup r. On the basis of results for the second stage test, distortion in the total group r was shown to be a function of the pattern of subgroup mean differences, total group sample size, and the magnitude of the pooled within-subgroup r. Implications were discussed. (Author)

Session 28.04

THE EFFECT OF POOLING TWO HETEROGENEOUS SUBGROUPS ON THE PRODUCT-MOMENT CORRELATION COEFFICIENT[1]

Alan L. Sockloff

Temple University

An equation was derived to determine the relationship between the pooled within-subgroup r and the r obtained from the total group data. It was, thus, possible to assess the amount of distortion introduced by pooling heterogeneous subgroups. As a basis for deciding whether to pool two subgroups in order to calculate a single r for the total group, a two-stage procedure was recommended: (1) comparison of the two within-subgroup r's; and (2) comparison of the total group r and the pooled within-subgroup r. On the basis of results for the second stage test, distortion in the total group r was shown to be a function of the pattern of subgroup mean differences, total group sample size, and the magnitude of the pooled within-subgroup r. Implications were discussed.

[1]Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974. Portions of this paper will be published in the Summer, 1975 issue of Educational and Psychological Measurement.

Frequently, in the psychological and educational literature, correlational studies are reported in which product-moment correlations are calculated between two variables for sets of data pooled across two, possibly heterogeneous, categorical subgroups. The pooling of heterogeneous subgroups to calculate a product-moment correlation was first discussed by Karl Pearson (Pearson, Lee, & Bramley-Moore, 1899). By way of illustration, these authors presented correlational data between length and breadth of skulls for 806 males (r = .0869) and 340 females (r = -.0424). When the two subgroups were pooled, an r of .1968 was obtained, and this r was considered to represent a large spurious correlation.

From a sampling of recent introductory statistics textbooks in psychology and education, it was found that writers do discuss the effects on the correlation coefficient resulting from pooling heterogeneous subgroups (Games & Klare, 1967; Glass & Stanley, 1970; Guilford, 1965; Walker & Lev, 1969; among others). Where references are made, Dunlap's (1937) paper on the combinative properties of correlation coefficients is most frequently cited. Dunlap presented a method for calculating a total group correlation coefficient from subgroup correlation coefficients and the means and standard deviations of the variables.

To date, a mathematical formulation of the effects on the correlation coefficient from pooling heterogeneous subgroups has been lacking. In lieu of such a formulation, textbook writers have tended to stress cautious interpretation and the use of subgroup correlation coefficients to help provide a rational explanation for correlation results in the total group. The major interest of this paper is the derivation of a mathematical formulation and the demonstration of the effects of

pooling two subgroups on the total group correlation coefficient. Of additional interest is a procedure to guide the decision concerning the pooling of data for the purpose of calculating a single correlation coefficient.

Formulation

Given two subgroups, let n_1 and n_2 be the subgroup sample sizes, where $n_1 + n_2 = N$. Let U and V be the distances between the subgroup means for X and Y, respectively, i.e., $U = \bar{X}_2 - \bar{X}_1$ and $V = \bar{Y}_2 - \bar{Y}_1$.

Sum of Squares and Cross Products

The sum of cross products for the total group, $SS(XY)_t$, can be defined

$$SS(XY)_t = SS(XY)_1 + SS(XY)_2 + UV\left(\frac{n_1 n_2}{N}\right), \qquad [1]$$

where $SS(XY)_1$ and $SS(XY)_2$ are the within-subgroup sums of cross products. Since the sum of squares for X in the total group is actually the sum of cross products with respect to itself,

$$SS(X)_t = SS(X)_1 + SS(X)_2 + U^2\left(\frac{n_1 n_2}{N}\right). \qquad [2]$$

Similarly, for Y,

$$SS(Y)_t = SS(Y)_1 + SS(Y)_2 + V^2\left(\frac{n_1 n_2}{N}\right). \qquad [3]$$

Total Group r_t

Using Equations 1, 2, and 3, the correlation coefficient for the total group is:

$$r_t = \frac{SS(XY)_1 + SS(XY)_2 + UV(n_1 n_2/N)}{\sqrt{\{SS(X)_1 + SS(X)_2 + U^2(n_1 n_2/N)\}\{SS(Y)_1 + SS(Y)_2 + V^2(n_1 n_2/N)\}}}.$$

An equivalent form of this equation was derived by Dunlap (1937). If $SS(XY)_W = SS(XY)_1 + SS(XY)_2$, $SS(X)_W = SS(X)_1 + SS(X)_2$, and $SS(Y)_W = SS(Y)_1 + SS(Y)_2$, then the equation defining r_t may be simplified:

$$r_t = \frac{SS(XY)_W + UV(n_1n_2/N)}{\sqrt{\{SS(X)_W + U^2(n_1n_2/N)\}\{SS(Y)_W + V^2(n_1n_2/N)\}}}. \quad [4]$$

The correlation coefficient for the total group is therefore expressed in terms of pooled within-subgroup sums of squares and cross products, subgroup sample sizes, and distances between the subgroup means.

Pooled Within-Subgroup r_W

The pooled within-subgroup correlation coefficient is obtained from pooling of the subgroup sums of squares and cross products:

$$r_W = \frac{SS(XY)_1 + SS(XY)_2}{\sqrt{\{SS(X)_1 + SS(X)_2\}\{SS(Y)_1 + SS(Y)_2\}}} = \frac{SS(XY)_W}{\sqrt{SS(X)_W SS(Y)_W}}. \quad [5]$$

It should be clear from inspection of Equation 5 that for equal variances of X in both subgroups and equal variances of Y in both subgroups, r_W is a weighted arithmetic mean of the within-subgroup correlations, weighted by the number of observations in each subgroup.

Furthermore, r_W may be compared to r_t in two ways. First, r_W is a special case of r_t resulting when subgroup mean differences are non-existent (i.e., $U = V = 0$). Second, r_W is that special case of r_t when subgroup differences are eliminated statistically. The latter comparison requires the form of a first-order partial correlation $r_{X_tY_t \cdot Z}$, where X_t and Y_t are the two variables measured in the total group and Z is a dichotomous variable indicating subgroup membership. In the formula for the first-order partial correlation, since r_{X_tZ} and r_{Y_tZ} are point-biserial correlations, the result of operating upon the formula for the first-order partial correlation is:

$$r_{X_tY_t \cdot Z} = \frac{SS(XY)_t - UV(n_1n_2/N)}{\sqrt{\{SS(X)_t - U^2(n_1n_2/N)\}\{SS(Y)_t - V^2(n_1n_2/N)\}}}.$$

Since the above equation represents an alternative definition of r_W, then r_W, the pooled within-subgroup correlation coefficient, is also

the result of statistically eliminating subgroup differences from the total group correlation coefficient. The latter definition of r_w suggests that r_w can be meaningfully used as a descriptive statistic with a known sampling distribution.

Further Derivation of r_t

If the numerator and denominator of Equation 4 are each divided by the product of the pooled within-subgroup standard errors of the mean ($s_{\bar{X}_w}$ and $s_{\bar{Y}_w}$), a more convenient definition of r_t arises. To complete this series of operations, by defining $t_X = U/s_{\bar{X}_w}$ and $t_Y = V/s_{\bar{Y}_w}$ as subgroup mean differences measured in units of standard errors, the following final form results:

$$r_t = \frac{(N - 2)r_w + t_X t_Y}{\sqrt{\{(N - 2) + t_X^2\}\{(N - 2) + t_Y^2\}}}. \qquad [6]$$

In this form, r_t is defined in terms of total group sample size, the pooled within-subgroup correlation coefficient, and subgroup mean differences that are measured in units distributed as Student's t under the appropriate assumptions.

The Decision to Pool

The following is a simple two-stage procedure, recommended as a basis for the decision to pool two subgroups in order to calculate a single correlation coefficient for the total group data.

1. The two within-subgroup correlation coefficients should be compared. Under the hypothesis H_0: $\rho_1 = \rho_2$, the unit-normal z test, relying on Fisher's r-z transformation (z_f), can be used:

$$z = \frac{z_{f_1} - z_{f_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}. \qquad [7]$$

If H_0 is rejected, it is unreasonable to calculate r_w as a measure of

correlation for the two subgroups. If H_o is accepted, then r_w can be considered a useful measure of correlation for the two subgroups, and the second stage should be followed.

2. In order to assess the distortion introduced by pooling the two subgroups, r_t should be compared to r_w under the hypothesis H_o: $\rho_t = \rho_w$. For this test, if r_w is considered an estimate of ρ_w,

$$z = \frac{z_{f_t} - z_{f_w}}{\sqrt{\frac{1}{N - 4}}} \qquad [8]$$

If H_o is rejected, then r_t may be considered distorted, but r_w can be used as a measure of correlation for the two subgroups. If H_o is accepted, then pooling the data from the two subgroups in order to calculate a total group correlation coefficient appears to be a parsimonious and reasonable procedure.

Examples of Distortion

According to Equation 6, distortion in r_t is affected by subgroup centroid differences and total group sample size. Three patterns of subgroup centroid differences are of interest and are shown in Figure 1.

Insert Figure 1 about here

---

(1) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on both variables (Cases a, b, and c of Figure 1), the greater the difference between the subgroups on the two variables, the more exaggerated the value of r_t in a positive direction. (2) If the mean of Subgroup 2 is higher than the mean of Subgroup 1 on one variable, and equal on the other variable (Cases d, e, and f of Figure 1), the greater the difference between the two subgroups on the one variable, the closer r_t is to a value of 0.00. (3) If the mean of Subgroup 2 is higher than

the mean of Subgroup 1 on one variable, and lower on the other variable (Cases g, h, and i of Figure 1), the greater the difference between the subgroups on the two variables, the more exaggerated the value of r_t in a negative direction. Furthermore, for constant differences between the subgroup centroids as measured in standard errors, increasing the total group sample size serves to minimize the effects of subgroup centroid differences, i.e., r_t approaches r_w.

The effects of subgroup sample size discrepancy on the calculation of r_t and r_w can be shown by reference to Equations 4 and 5. According to Equation 5, for constant total group sample size, the larger the discrepancy between the subgroup sample sizes, the greater the influence of the larger subgroup in the calculation of r_w. In addition, according to Equation 4, for constant differences between subgroup centroids and for constant total group sample size, the larger the discrepancy between n_1 and n_2, the smaller the effect of subgroup centroid differences in the calculation of r_t, and, thus, the more equal the values of r_t and r_w.

For the second stage test, in order to demonstrate the amount of distortion introduced by pooling heterogeneous subgroups, Equation 6 was utilized to calculate r_t under varying sample conditions. The sample conditions were derived from combinations of four magnitudes of subgroup centroid differences for the three patterns, five total group sample sizes (N), and three values of r_w. By assuming two bivariate normal populations, the magnitude of difference between subgroup means can be represented by employing four-decimal critical values of Student's t distribution for $p<.05$, $p<.01$, $p<.001$, and $p<.0001$, obtained for N-2 df from Sockloff & Edney's (1972) tables. The .001 and .0001 significance levels were used because these levels represent extreme differences that are sometimes found in research data, although not necessarily

reported. The five total group sample sizes were 10, 50, 100, 200, and 1000. The three values of r_w were .8000, .4000, and 0.0000, chosen to represent high, moderate, and low correlations, respectively.

Tables 1, 2, and 3 present calculated values of r_t derived from the values of r_w under the varying sample conditions. Also included in these tables are the second stage two-tailed tests of H_0: $\rho_t = \rho_w$ to assess the amount of distortion introduced under the conditions. Negative values of r_w are not shown in the tables since the effects of the three patterns for negative r_w's are opposite in sign from those shown for positive r_w's, e.g., the amount of exaggeration in a positive direction for a positive correlation under Pattern 1 is equal to the amount of exaggeration in a negative direction for a negative correlation under Pattern 3.

As shown in Table 1, under Patterns 1 and 2, large differences

---

Insert Table 1 about here

---

between the subgroup means have a small effect on the value of r_t when $r_w = .8000$. Under both patterns, a total group sample size of 50 appears to be sufficient to minimize the distortion introduced by subgroup mean differences that are significant at the .0001 level. On the other hand, the results were quite different under Pattern 3. When the means of the two subgroups are significantly different at the .0001 level, but in opposite directions, for a total group sample size of 50 r_t was calculated to be .3089, which is significantly different from an r_w of .8000. Furthermore, even for a total group sample size of 1000, the pooling of subgroups when subgroup means differ in opposite directions at the .0001 level produced an r_t of .7729. Although this value of r_t is significantly different from an r_w at the .05 level,

one can argue that such statistically significant differences between r_t and r_w have little practical significance.

According to Table 2, when $r_w = .4000$, the results for Pattern 1

Insert Table 2 about here

suggest that total group sample sizes of 50 are sufficient to avoid distortions introduced by pooling subgroups when both sets of subgroup means differ significantly in the same direction at the .0001 level. Under Pattern 2, significant distortion was not found, even for a total group sample size of 10. The Pattern 3 results suggest that a total group sample size of 200 will avoid distortion in the calculation of r_t.

According to Table 3, total group sample sizes of 50 appear to be

Insert Table 3 about here

sufficient to avoid distortion when $r_w = 0.0000$ and the two sets of subgroup means differ at the .0001 level. Based on the symmetry of the sampling distributions of r when $\rho = 0$, this conclusion holds for subgroup means differing in the same or opposite directions.

Discussion

The various results clearly suggest the varieties of distortion that may be introduced by haphazardly pooling subgroups of data for the purpose of calculating a single correlation coefficient. The two-stage test procedure should offer protection against such distortions. In addition, it was shown that greater latitude exists in terms of non-distorting pooling when the subgroup mean differences are small, the subgroup sample sizes are large, and the pooled within-subgroup correlations are low to moderate. The calculated examples suggest limits within which distortion does not seriously affect correlational results.

The types of subgroups to which this discussion refers are those resulting from natural dichotomies and those resulting from an arbitrary split where (a) the decision to split the total group was based on considerations other than that of ridding the data of non-linearity, and (b) middle range data has been discarded. When an arbitrary split is made to rid the total group data of non-linearity, the two subgroups may show evidence of different, but linear, relationships. According to the first stage test, if the two within-subgroup correlations are different, then it would appear unreasonable to even consider pooling the data on the basis of the original rationale for having made the split. On the other hand, if middle range data is not discarded when an arbitrary split is made for reasons other than ridding the data of non-linearity, then pooling would appear to be a reasonable step toward restoring the information contained within the total bivariate set of data.

Implications of this study relate to the use of the pooled within-subgroup correlation coefficient, to current practices in educational research, and to generalizations of this study in terms of pooling multiple subgroups in the calculation of correlation matrices. First, assuming no difference between the within-subgroup correlation coefficients (non-rejection of the first stage test), the pooled within-subgroup correlation coefficient is useful as a descriptive statistic with hypothesis-testing capabilities resulting from its equivalence to a first-order partial correlation coefficient. Second, in research involving multiple dependent measures that are analyzed via several ANOVA's, rather than MANOVA, intercorrelations among the dependent measures should be assessed through pooled within-subgroup correlation coefficients rather than total group correlation coefficients. Otherwise,

the meaningfulness of the intercorrelations would be contingent upon the failure to find subgroup differences in all of the ANOVA's.

Last, considering the demonstrated varieties of possible distortion of a single correlation coefficient from the haphazard pooling of only two heterogeneous subgroups, the generalizations of these results must be inherently more complex, i.e., the effects of pooling multiple subgroups on correlation matrices. If, indeed, such complex distortion can be demonstrated, and it is desirable to pool data for reasons of parsimony, this suggests that further study should be devoted to the multivariate case and the development of appropriate test procedures.

References

Dunlap, J. W. Combinative properties of correlation coefficients. Journal of Experimental Education, 1937, 5, 286-288.

Games, P. A., & Klare, G. R. Elementary statistics: Data analysis for the behavioral sciences. New York: McGraw-Hill, 1967.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N. J.: Prentice-Hall, 1970.

Guilford, J. P. Fundamental statistics in psychology and education. (4th ed.) New York: McGraw-Hill, 1965.

Pearson, K., Lee, A., & Bramley-Moore, L. B. Genetic (reproductive) selection: Inheritance of fertility in man and of fecundity in thoroughbred racehorses. Philosophical Transactions of the Royal Society (Series A), 1899, 192, 257-330.

Sockloff, A. L., & Edney, J. N. Some extensions of Student's t and Pearson's r central distributions. Technical Report 72-5. Philadelphia: Temple University, 1972.

Walker, H. M., & Lev, J. Elementary statistical methods. (3rd ed.) New York: Holt, Rinehart, & Winston, 1969.

Table 1

Values of r_t Resulting from Three Patterns of Subgroup Centroid Differences and Five Total Group Sample Sizes: r_w = .8000

Total group sample size	Significance levels for t distribution when subgroup mean differences equal critical values			
	$p<.05$	$p<.01$	$p<.001$	$p<.0001$
Pattern 1: $\bar{X}_2 > \bar{X}_1$, $\bar{Y}_2 > \bar{Y}_1$, both significant				
10	.8799	.9169	.9521	.9727*
50	.8155	.8261	.8408	.8546
100	.8077	.8132	.8210	.8288
200	.8039	.8066	.8107	.8148
1000	.8008	.8013	.8022	.8030
Pattern 2: $\bar{X}_2 > \bar{X}_1$, significant; $\bar{Y}_1 = \bar{Y}_2$				
10	.6200	.5156	.3914	.2953
50	.7683	.7460	.7138	.6822
100	.7844	.7732	.7568	.7403
200	.7923	.7867	.7784	.7699
1000	.7985	.7973	.7957	.7940
Pattern 3: $\bar{X}_2 > \bar{X}_1$, $\bar{Y}_2 < \bar{Y}_1$, both significant				
10	.0813*	-.2523***	-.5691****	-.7547****
50	.6602*	.5654**	.4332****	.3089****
100	.7305	.6816**	.6108***	.5412****
200	.7653	.7405*	.7040**	.6672****
1000	.7931	.7881	.7806	.7729*

Note.--Asterisks refer to significance levels of unit-normal z tests comparing r_t and r_w.

* $p<.05$

** $p<.01$

*** $p<.001$

**** $p<.0001$

Table 2

Values of r_t Resulting from Three Patterns of Subgroup Centroid Differences and Five Total Group Sample Sizes: r_w = .4000

Total group sample size	Significance levels for t distribution when subgroup mean differences equal critical values			
	$p<.05$	$p<.01$	$p<.001$	$p<.0001$
Pattern 1: $\bar{X}_2 > \bar{X}_1$, $\bar{Y}_2 > \bar{Y}_1$, bot significant				
10	.6396	.7508	.8564	.9182**
50	.4466	.4782	.5223	.5637
100	.4232	.4395	.4631	.4863
200	.4116	.4198	.4320	.4443
1000	.4023	.4040	.4065	.4090
Pattern 2: $\bar{X}_2 > \bar{X}_1$, significant; $\bar{Y}_1 = \bar{Y}_2$				
10	.3100	.2578	.1957	.1477
50	.3842	.3730	.3569	.3411
100	.3922	.3866	.3784	.3701
200	.3961	.3933	.3892	.3850
1000	.3992	.3987	.3978	.3970
Pattern 3: $\bar{X}_2 > \bar{X}_1$, $\bar{Y}_2 < \bar{Y}_1$, both significant				
10	-.1590	-.4184*	-.6648**	-.8092***
50	.2913	.2175	.1147*	.0181**
100	.3459	.3079	.2529	.1987*
200	.3730	.3537	.3253	.2967
1000	.3946	.3907	.3849	.3789

Note.--Asterisks refer to significance levels of unit-normal z tests comparing r_t and r_w.

* $p<.05$

** $p<.01$

*** $p<.001$

Table 3

Values of r_t Resulting from One Pattern of Subgroup Centroid Differences and Five Total Group Sample Sizes: $r_w = 0.0000$

Total group sample size	Significance levels for t distribution when subgroup mean differences equal critical values			
	$p<.05$	$p<.01$	$p<.001$	$p<.0001$
Pattern 1: $\bar{X}_2 > \bar{X}_1$, $\bar{Y}_2 > \bar{Y}_1$, both significant				
10	.3993	.5846	.7606*	.8637**
50	.0777	.1303	.2038	.2728
100	.0386	.0658	.1051	.1438
200	.0193	.0330	.0533	.0738
1000	.0038	.0066	.0108	.0151

Note.--Asterisks refer to significance levels of unit-normal z tests comparing r_t and r_w.

* $p<.05$

** $p<.01$

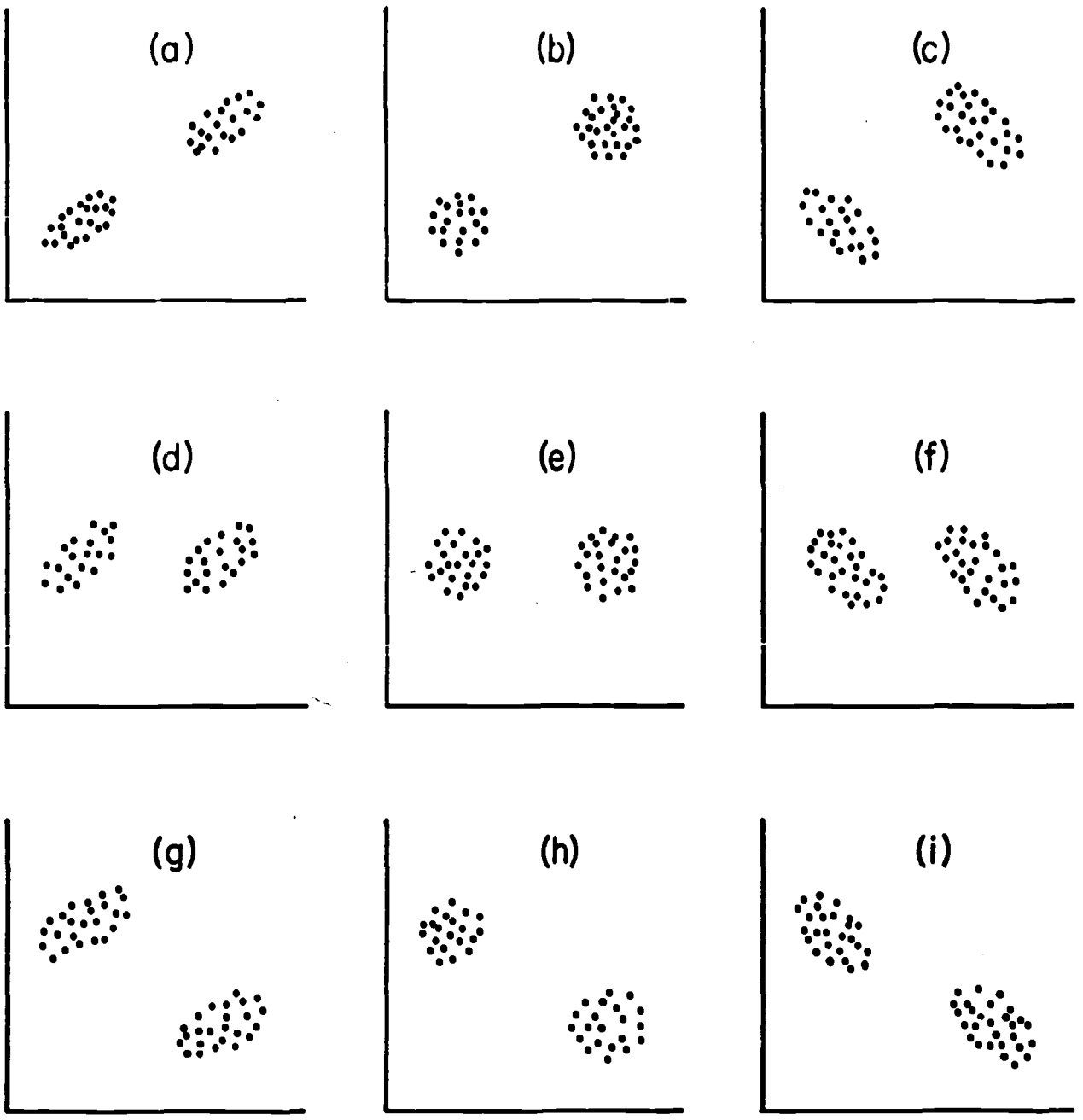

FIG. 1. Nine exaggerated bivariate plots of total group data resulting from pooling heterogeneous subgroups.